Normative foundations of human cooperation

Fehr, Ernst ; Schurtenberger, Ivo

Abstract: A large literature shares the view that social norms shape human cooperation, but without a clean empirical identification of the relevant norms almost every behaviour can be rationalized as norm driven, thus rendering norms useless as an explanatory construct. This raises the question of whether social norms are indeed causal drivers of behaviour and can convincingly explain major cooperation-related regularities. Here, we show that the norm of conditional cooperation provides such an explanation, that powerful methods for its empirical identification exist and that social norms have causal effects. Norm compliance rests on fundamental human motives ('social preferences') that also imply a willingness to punish free-riders, but normative constraints on peer punishment are important for its effectiveness and welfare properties. If given the chance, a large majority of people favour the imposition of such constraints through the migration to institutional environments that enable the normative guidance of cooperation and norm enforcement behaviours.

DOI: https://doi.org/10.1038/s41562-018-0385-5

A slightly shorter version of this paper has been published in NATURE HUMAN BEHAVIOR July 2018, pp 458-468

Normative Foundations of Human Cooperation

Ernst Fehr & Ivo Schurtenberger

A large literature shares the view that social norms shape human cooperation. However, in the absence of a clean empirical identification of the relevant norms almost every behavior can be rationalized as norm-driven, thus rendering social norms useless as an explanatory construct. This raises the question whether there is a parsimonious and empirically convincing social norms-based explanation for major cooperation-related behavioral regularities and whether norms are indeed causal drivers of cooperative behavior. Here, we show that such an explanation is possible, that various powerful methods for the empirical identification of social norms exist, that they all support the existence of a norm of conditional cooperation and that social norms have causal effects. Norm compliance rests on fundamental motivational inclinations of humans ("social preferences") that also imply a willingness to punish free-riders but normative constraints on peer punishment are important for its effectiveness and welfare properties. If given the chance, a large majority of people favor the imposition of such constraints through the migration to institutional environments that enable the normative guidance of cooperation and norm enforcement behaviors.

Normative constraints and prescriptions are ubiquitous and pervade almost every aspect of human social life, from the mundane to the most profound. They appear to play a role in all social groups and they have been documented for a large number of ancient societies[1,2] but also play a role in contemporary societies. Norms are part of the weave of social life and, if obeyed, they make it predictable, constitute social order and become the cement of society[3] but if compliance with fundamental norms breaks down – as it sometimes happens in the aftermath of lost wars or natural disasters – disorder, revolt or revolutionary chaos prevails, and life becomes "solitary, poor, nasty, brutish and short".[4]

Human cooperation is an equally ubiquitous phenomenon that is present in some form in almost every social relationship and is key for the success of social units from the family to the nation state to global organizations[5]. Sometimes, cooperation is in the material self-interest of people but here we are interested in those aspects of cooperation where economic incentives alone are not sufficient to induce individuals to cooperate because free-riding would maximize their private gains but collectively the group would be better off if it's members cooperated. Throughout human history, myriad scenarios are characterized by such social dilemmas. Every successful sequential exchange, in which

one party provides the quid pro quo first, constitutes an act of cooperation because the first-moving party has to trust while the second-moving party has to refrain from cheating. Our ancestors also faced social dilemmas when they hunted large game, during tribal warfare or reciprocal food sharing in times of need. Contemporary humans encounter them in team production settings and whenever there is a tension between one's own interest and the reputation of the company, when paying taxes despite low probabilities of being caught in tax evasion or in the context of problems of a truly global scale such as climate change.

To what extent, and how do social norms shape human cooperation? There are social norms such as the norm to keep a promise or the honesty norm that affect behavior in cooperative contexts although the behavioral prescriptions of these norms are not directly related to cooperation. For example, the honesty norm proscribes lying and that implies that one should also not lie to evade taxes and the norm to keep one's promises implies that one should also keep promises made to an exchange partner but these norms have implications that go far beyond cooperative contexts. In this review, we focus instead on social norms that *directly* prescribe, and limit their prescription to, cooperation and punishment behaviors in social dilemma and collective action contexts. An example of such a norm is the "conditional cooperation norm" which we define in more detail below. We ask whether these norms can, in principle, explain major behavioral regularities observed in collective action contexts, what the properties of these norms are and which motivational forces ensure compliance with them, and whether they indeed guide or are the causal drivers of behavior in collective action.

To answer these questions requires a clear definition of social norms. We define them as *commonly known standards of behavior that are based on widely shared views how individual group members ought to behave in a given situation*[3,6,7]. This definition entails three crucial features of social norms. First, a social norm establishes a normative standard of behavior that applies to a particular group and to a particular situation. Second, the norm is not defined in terms of group members' actual behavior nor in terms of their motives, their compliance or the conditions under which compliance occurs; it is exclusively defined in terms of a normative behavioral standard, i.e., how group members *ought to* behave. Third, this normative standard and its widely shared approval is commonly known by group members.

Because a norm requires that the normative standard is widely shared, non-compliance with the norm automatically triggers some disapproval. Therefore, if individuals dislike the thought that others disapprove of them they automatically have some incentive to comply although, as we will see, this incentive may not necessarily be sufficient to induce compliance. We will therefore also ask the question which kind of other motives and mechanisms support compliance with social cooperation norms and whether they act as a constraint on potential non-compliers or are part of the "intrinsic"

motivation of individuals. In this context we will also ask whether the (peer) punishment of norm violators is itself a social norm or whether it is driven by other motivational sources.

The above definition of a social norm implies that norms are not just a property of an individual – they constitute a property of the group and are, therefore, collective phenomena that may shape individuals' behaviors, expectations and sometimes even their deep individual properties, i.e., their intrinsic motives and preferences. If norms – for example, norms of equity and reciprocity – also shape individuals' preferences, individuals must themselves be regarded as partly "constituted" through social practices. In other words, society also exists "within" the individuals.

Fundamental regularities in cooperation-related behaviors?

To assess the role of social norms for human cooperation, we describe in a first step major behavioral regularities observed in experimental social dilemma games. With the exception of experiments that allow for face-to-face communication, the subjects in these games are anonymous to each other. They play for real money under conditions where complete free-riding is the dominant strategy for selfish individuals in one-shot games and backward induction implies that complete free-riding is also predicted in the finitely repeated game. We deliberately restrict ourselves to these experimental settings because to precisely identify the role of social norms their predictions must differ from the self-interest model. Field evidence, in contrast, typically does not allow to rule out self-interest with perfect certainty but below we will point out that many lab observations resemble regularities that are observed in naturally occurring environments. In a second step, we then discuss the ability of social norms to provide a parsimonious explanation for the regularities.

The following patterns are among the key findings in the literature:

(1) Although complete free-riding is a dominant strategy, a substantial share of the subjects cooperate in one-shot social dilemmas but free-riding frequently also prevails[8,9]. However, if subjects can communicate about the game before they play it cooperation strongly increases[8,10,11] (Fig. 1a).

(2) A large proportion of subjects are conditional cooperators, that is, the belief that other group members cooperate at high levels induces them to also cooperate at high levels but if others are believed to decrease their cooperation these individuals also decrease their cooperation[12-14] (Fig. 1b).

(3) In finitely repeated public good games (PGG), cooperation is initially relatively high but often declines to very low levels towards the final periods[15,16]. This holds regardless of whether the game is frame as a public goods game or as a common pool resource game[17]. If subjects play the

finitely repeated game several times – but each time with a new composition of group members – cooperation always starts high and becomes very low towards the end of the game[18] (Fig. 1c).

(4) In finitely repeated public good games, cooperation is generally higher in groups with a stable group composition ("partner matching") compared to random reassignment of individuals to groups in every period ("stranger matching")[14,18,19] (Fig. 1d).

(5) Merely framing a simultaneously played prisoners' dilemma game differently by calling it Community Game instead of Stock Market Game typically causes substantial increases in cooperation rates. However, if the game is played sequentially this framing effect vanishes [20,21] (Fig 1e).

(6) There is a widespread willingness to punish free-riders even in one-shot interactions although it is costly for the punisher[22-24] (Fig. 1f). Furthermore, peer punishment opportunities in repeated interactions cause large cooperation increases and often lead to near complete and stable cooperation under partner matching[22,25] (Fig. 1i). These opportunities are, however, also associated with high initial costs such that group welfare does not increase (or even decreases) for roughly 10 periods[22,25].

(7) The effectiveness of peer punishment in enhancing cooperation is undermined if punishment threats signals selfish intentions[26-29] and by "perverse"[30] or "antisocial" punishment[31,32] of cooperators in public good games by those who free-ride – a tendency that varies strongly across different cultures (Fig. 1g).

(8) Despite the high initial cost caused by peer-punishment, subjects eventually prefer environments with a peer punishment opportunity almost unanimously over an environment that rules out peer punishment[33,34] (Fig. 1 h).

(9) The opportunity to reward cooperators – either through the preferred choices of cooperative partners[35] or through the direct rewarding of those with a high reputation for cooperation[36-39] causes large cooperation increases (Fig. 1i).

(10) Stable cooperation at very high levels can be achieved when (i) cooperative individuals are exogenously matched together[40,41] (Fig. 1j) or (ii) in intergenerational public good games when individuals can give advice that is common knowledge to the next generation[42].

An important question is how insights gained in lab experiments relate to behavior in naturally occurring environments. Several studies[43-52] demonstrate that individuals' behavior in the lab is predictive of their behavior in relevant field settings. For instance, people who tend to contribute more in public good games are more likely to participate in local and national accountability institutions[43]. Fishermen who exhibit more cooperation in a laboratory public good game also show more cooperative behavior in a real world common pool resource problem by employing more sustainable

fishing techniques; they use buckets with larger holes such that younger shrimps are not yet caught[44]. One study[45] shows that Ethiopian communities that face serious common pool resource problems are better able to maintain the commons if they have a higher share of people that display conditional cooperation in a public goods experiment. This study also provides evidence suggesting that causality runs from conditional cooperation to better maintenance of the commons resource. Behaviors consistent with conditional cooperation are also observed in field experiments[46]. Experimentally induced signs of uncooperative behavior, such as graffiti or unreturned shopping carts induces people to generate more disorder in public spaces. Another study[47] examines how punishment behavior by community leaders in a social dilemma experiment predicts the success these leaders have in managing the communities' forest commons. Leaders who punish antisocially see worse forest outcomes than those who emphasize efficiency and equality in their experimental punishment behavior. Finally, evidence from rural labor markets in India indicates the existence of strong collective action norms that prevent that casual daily laborers receive payments below the prevailing wage. Poor workers typically reject low paying job offers although this means that they remain unemployed for that day, and they are willing to incur personal costs to punish workers who accepts wage cuts, even when such a norm violators are strangers that come from an unrelated labor market[48]. As a consequence, nominal wages exhibit substantial downwards rigidity that generates substantial reductions in labor demand[49].

Can social norms explain the regularities in cooperation-related behaviors?

All above mentioned regularities are largely incompatible with the pure self-interest model, i.e., they cannot be explained if it is common knowledge that all actors are rational and selfish. When free-riding is the selfishly dominant strategy these actors will never cooperate in a (subgame perfect) equilibrium of the public goods game, regardless of whether they are in a one-shot encounter or a finitely repeated game with partner or stranger matching. Likewise, because punishment/reward is costly for the punisher/rewarder subjects will never punish or reward in a (subgame perfect) equilibrium and, therefore, punishment/reward opportunities are futile. This also means that there is no reason for subjects to prefer an environment with peer punishment opportunities. Finally, it is also not possible to generate cooperation through communication, advice giving or the sorting of individuals if selfishness and rationality are common knowledge.

However, many of these regularities can, at least in principle, be explained if one directly assumes that a significant share of individuals has a desire to comply with a social cooperation norm[53]. We call this the direct social norms approach[7,41,54,55] because it directly assumes (i) the existence of a norm c^* that is defined in terms of a specific behavior and (ii) that individuals have an intrinsic desire to comply with

c^* without providing a deeper micro-foundation of c^* and motives for norm compliance. In the context of cooperation, c^* describes the smallest cooperation level that is consistent with the normative prescription. Formally, this can be modelled by a utility function u_i in which individual i's utility depends positively on i's own material payoff x_i (which depends on all players' choices) while negative deviations of i's behavior c_i from the social norm c^* ($\mathrm{c}_i < c^*$) generate some disutility:

$$u_i = \begin{cases} \mathrm{x}_i - \gamma_i(\mathrm{c}_i - c^*)^2 & if\ \mathrm{c}_i < c^* \\ \mathrm{x}_i & if\ \mathrm{c}_i \geq c^* \end{cases}$$

The term $\gamma_i(\mathrm{c}_i - c^*)^2$ denotes the psychic cost of deviating from the social norm (for simplicity these costs increase quadratically with negative deviations from the norm ($\mathrm{c}_i - c^*{}_\mathrm{i}$) and $\gamma_i \geq 0$ captures an individual's strength of the desire to conform to the norm. This approach represents a simple theory of conformism based on the assumption that negative deviations from the norm are, for some reason, psychologically costly for individuals with a strictly positive γ_i. In the context of cooperation, higher individual cooperation levels c_i are costly and thus reduce the individual's material payoff x_i but if c_i is below the norm c^* an increase in c_i reduces the costs of non-conformity $\gamma_\mathrm{i}(\mathrm{c}_\mathrm{i} - \mathrm{c}^*)^2$. For a sufficiently large level of γ_i the individual has therefore an incentive to obey the social norm c^*. Note that we assume for simplicity that positive deviations from the norm c^* have no psychological costs or benefits.

It is almost surely the case that the psychological cost of negative deviations from c^* (i.e., the γ_i's) vary across people but the assumption that there are some psychological costs of negative deviations makes sense in the light of the definition of a social norm because that definition implies that group members widely approve of the norm and that this is known by the subjects. Thus, subjects know that if they violate a social norm they are likely to face the disapproval of other people and for some people even the mere thought that others might disapprove of their action could constitute a psychological cost. In principle, γ_i could also represent the cost of deviating from a behavioral habit acquired in social life. Or the psychological cost of noncompliance could positively depend on the how widely the norm is shared among the group members. However, in the following we assume for simplicity that γ_I is fixed and varies across individuals.

Unconditional normative prescriptions like "be selfless", "do the right thing" or "be moral" cannot explain the behavioral regularities described above. For example, they can neither explain communication effects (fact 1) nor can they explain the decline in cooperation over time (fact 2) or the higher levels of cooperation in partner compared to stranger matching (fact 3). In contrast, a social norm of *conditional* cooperation can help explain all regularities but those described in fact (6-8). This norm prescribes full cooperation as long as other group members also cooperate fully but if others' average cooperation becomes smaller it is normatively justified to match this reduction, that is, the

conditional cooperation norm prescribes to contribute at least as much as others' average contribution. Note that this implies that subjects' empirical beliefs about others' average cooperation become an important determinant of their cooperation levels – the more others cooperate the higher is the incentive to cooperate for an individual with a positive γ_i. This explains fact (2).

But this norm can also explain regularity (1): subjects with a very small γ_i ($\gamma_i \approx 0$) will defect while those with a sufficiently large γ_i and a high expectation about others' cooperation will cooperate in one-shot social dilemmas. Moreover, under face-to-face communication subjects often promise to each other to cooperate[56] which is very likely to increase beliefs about others' cooperation. This increase in others' expected cooperation will then induce individuals with a sufficiently positive γ_i to increase their cooperation levels.

It has been shown[5,57] that the existence of *imperfect* conditional cooperators is the key ingredient for explaining fact (3) – the decay of cooperation over time in finitely repeated games. Conditional cooperation is imperfect if an individual does not match other group member's average cooperation perfectly but cooperates somewhat less than others are expected to cooperate on average. The above utility function assumes that people care positively for their own payoff and, therefore, individuals with a positive yet sufficiently low γ_i will not obey the norm c^* perfectly but reduce c_i somewhat below c^*, which implies imperfect conditional cooperation. However, if many individuals cooperate less than what each of them expect others' to cooperate, jointly their expectations are too optimistic, which results in a downwards revision of their expectations and this then leads – via conditional cooperation – to a further decline in their cooperation rates, etc., etc..

The existence of a conditional cooperation norm can also explain fact 4 – the higher cooperation rates under a stable group composition – and fact 5, the existence of a framing effect on cooperation in the simultaneously played PD but not in the sequentially played PD[20]. When there is a stable group composition, even selfish individuals (i.e., those with $\gamma_i \approx 0$) have temporarily a strong incentive to cooperate because this generates benefits in future periods by inducing conditional cooperators to keep contributing (fact 4). To explain fact 5, recall that if there is a norm of conditional cooperation subjects who derive disutility from norm violations adjust their cooperation level to what they believe the other player will do in the simultaneously played PD. For optimistic beliefs they cooperate, for pessimistic beliefs, they defect. Under the plausible assumption that the label "Community Game" renders beliefs about the partner's cooperation more optimistic, conditionally cooperative subjects will cooperate with higher frequency. However, for the second mover in the sequential PD beliefs are irrelevant because this player already knows exactly what the first-mover did. Thus, the frame can no longer change beliefs and thus becomes irrelevant; and if a rational first mover anticipates the absence of a framing effect (s)he has no reason to condition behavior on the frame either. Note that this

explanation does not assume that the conditional cooperation norm changes across frames or between simultaneous and sequential play; a given norm can explain the changes in behavior.

The conditional cooperation norm can also explain why the addition of mutual reward opportunities to a public goods game increases cooperation (fact 9). In the presence of mutual reward opportunities subjects can observe the cooperation level of other group members in the public good game after which they can spend money on rewarding other group members that costs them less than it benefits the rewarded subjects. This basically boils down to the opportunity of playing another bilateral prisoners' dilemma (PD) with each of the other group members after they observed others' cooperation levels. Obviously, the norm of conditional cooperation also applies to these PD games and because cooperation in the public good game can serve as a signal of cooperative intent, cooperation in the public good game fosters the belief that an individual will also cooperate in the PD. Therefore, mutual reward opportunities increase the incentive to cooperate in the public goods game.

Finally, the conditional cooperation norm can also help explain fact (10), i.e., why the assignment of cooperative individuals to the same group may cause *high and stable* cooperation. In terms of the direct social norms approach, cooperative individuals may be viewed as those with a sufficiently high γ_i such that for them perfect obedience with the norm ($c_i = c^*$) becomes optimal. If, in addition, these subjects are told that they are grouped together with other cooperators[40] they start with high expectations that trigger high cooperation which confirms the initial high expectation. The publicly known sorting of cooperative individuals into a group thus renders cooperation an equilibrium outcome.

For a similar reason, the existence of a conditional cooperation norm may also explain why cooperative advice by a previous generation of players that is made common knowledge among all current group members (fact 10) causes large increases in cooperation rates. Cooperative advice that is common knowledge induces a general increase in the expected cooperation of other group members[42]. Together with the norm of conditional cooperation the increased expectations then give rise to a general increase in cooperation rates.

However, although a conditional cooperation norm can potentially explain many above-mentioned facts this does not mean that this norm is in fact the driver behind subjects' choices because there could be other motives – such as equity or reciprocity motives – that make similar predictions. Moreover, a norm of conditional cooperation cannot explain why subjects punish free-riders (fact 6) and thus also not subjects' preferences for playing the public goods game in an environment that allows for peer punishment (fact 8). This follows simply from the fact that the conditional cooperation norm is defined in the space of cooperation behavior and not in the space of punishment behavior. One may, of course, stipulate the existence of another norm that renders punishment of free-riders a

socially desirable act but (i) there is little evidence for this and (ii) it shows one of the drawbacks of an unconstrained direct social norms approach. By stipulating that a particular behavior constitutes a social norm it is possible to explain any behavior which renders such an approach irrefutable and thus empty – a problem that we take up later.

With respect to a potential norm that prescribes the punishment of free-riders, there is some evidence that punishers can acquire a reputation gain[58,59] if punishment of selfish behavior is the only way to acquire a prosocial reputation but if there are other ways (e.g., through altruistic helping) punishment is much less used as a signal of prosociality. This suggests that punishment is an inferior way of signaling prosociality. In addition, it has been found[60] that people do not support or reward those who punish free-riders and they do not approve of punishers more than they approve of non-punishers. This is in line with other evidence[61] that shows that those who punish free-riders the most even spend money to prevent that their punishment acts become public information, suggesting that they believe it might hurt their reputation. Moreover, even if the punishment of free-riders conferred a prosocial reputation and thus rendered the individual a desirable exchange partner, this would not indicate that punishment is a social norm. There are many altruistic behaviors (e.g., giving to the victims of a hurricane) that confer a prosocial reputation if observed by others but nevertheless are not normatively required behaviors.

With regard to the peer punishment of free-riders in lab experiment it is useful to relate them to punishment in the field. In reality, peer punishment ranges from a raised eye brow to a mild, yet hurtful, smile or verbal insult, from social ridicule to ostracism, from shaming on the internet to the expulsion from social groups. These forms of punishment differ widely in the cost for the punisher and their hurtfulness for the punished. But to the extent to which they are psychologically or economically costly for the punisher and impose psychological or economic cost on the punished they share the key features of punishment in the lab. Moreover, one important lesson from lab experiments is that even if punishment is costly and yields no material benefit for the punisher it nevertheless systematically occurs.

The psychology of norm compliance

To make progress in understanding the potential impact of social norms on human cooperation it is important to examine more closely the psychological reasons that induce individuals to comply with social norms. The direct social norms approach stipulates a normative behavioral standard and a psychological cost of non-compliance but does not provide a microfoundation for the behavioral standard and is typically not very explicit about the psychological cost of non-compliance. In principle, these costs could arise because individuals may be averse to actual, anticipated or merely imagined disapproval when deviating from the norm. In this case, compliance rests on an internalized desire for

conformism which has been challenged long ago as a general and sufficient basis for norm compliance[62].

Another reason for psychological costs of norm compliance arises if individuals have an intrinsic desire for equity or fairness and social norms play a role in defining what is perceived as equitable or fair[63-65]. This case is also methodologically interesting because it implies that a collective phenomenon – the social norm – substantively affects the content of individuals' motivation by influencing what is perceived as fair while the intrinsic desire for fairness then ensures compliance with the norm. A third reason for costs of deviating from the social norm could be that individuals have a desire to reciprocate the behavior of relevant others[66-68]. In this case, the reciprocity motive applies, i.e., the tendency to reward kind intentions with kindness ("positive reciprocity") and to punish hostile or unkind intentions ("negative reciprocity"). Note, however, that this motive requires a definition of what constitutes kind and unkind behavior which is typically also based on some normative notion of fairness/equity. For a reciprocally motivated individual psychic costs of non-compliance arise, if it fails to reciprocate to a kind act with kindness or does not retaliate to a hostile act with a hostile response. Therefore, as in the case of fairness/equity motives the reciprocity motive becomes operative on the basis of what is perceived as fair/kind and unfair/unkind.

A fourth reason for psychic costs of non-compliance arises if individuals have a propensity towards guilt aversion[69-71]. This theory rests on the idea that individuals experience the aversive, utility-decreasing, emotion of guilt if they disappoint others. A social norm only exists if group members widely approve of the norm, and if there is widespread compliance then an individual act of non-compliance is almost surely disappointing other individuals. For example, if a subject believes that her partner in the prisoners' dilemma expects her to cooperate then she disappoints him/her if she defects and if she feels guilt and anticipates this emotion she has an incentive to cooperate. Therefore, to the extent to which social norms generate the belief that others expect the individual to comply – a very likely belief in the presence of wide-spread compliance – a guilt averse individual has some incentive to cooperate. However, if a social norm is systematically violated such that the individual does not face a general expectation of compliance, a guilt averse individual has no reason to comply with the norm. Guilt aversion is thus likely to generate conditional norm compliance behavior that is mediated by individuals' beliefs about what others expect from them.

Finally, self-image theory assumes that individuals assign an intrinsic value to their self-image as a prosocial individual[72]. In this case, non-compliance with socially beneficial norms is detrimental for their self-image and provides a psychological deterrent for non-compliance. Similar to the case of fairness and reciprocity theories this approach rests on some pre-existing notion – the notion of "prosociality" – which is likely to be shaped by social norms.

It is interesting that all the above mentioned approaches rest on assumptions about *individuals' intrinsic* motivational properties. These motives – for example the desire for fairness – are assumed to be stable across contexts. Stability in the desire for fairness does not mean, however, that the *content* of what is defined as fair is stable across contexts. It only means that individuals' preferences for implementing what is defined as fair, i.e., their willingness to pay to implement the fair action, is stable while what is defined in a given society or group as fair or prosocial can be malleable. Thus, a main difference between social preference theories of equity, reciprocity, guilt aversion, and self-image and the direct social norms approach is that the former are concrete about the motivational basis of norm compliance and these motives are assumed to be stable across contexts whereas the direct social norms approach remains vague with respect to the motives underlying norm compliance.

For example, both conditionally cooperative behavior and the willingness to punish free-riders in a public goods game can arise from a desire for fairness or reciprocity. In other words, inequity averse subjects and reciprocity-motivated subjects are often conditional cooperators as well as punishers[63,68] and, therefore, these motives contribute to the explanation of all the major qualitative regularities mentioned above (except the existence of antisocial or perverse punishment which we discuss below). Likewise, the communication effects (fact 1) as well as the framing effects (fact 5) can be explained by *stable* preferences for equity or reciprocity because these preferences imply conditionally cooperative behavior such that if frames and pre-play communication renders expectations about others' cooperation more optimistic, subjects will cooperate more.

Or take, e.g., regularity (4) that "partners" generally cooperate more than "strangers". The theory of inequity aversion or reciprocity can explain this finding by the fact that the existence of inequity averse or reciprocal subjects generates incentives for selfish individuals in a partner treatment to invest into cooperation during the early periods of a finitely repeated game[18]. This investment is profitable because it maintains the cooperation of the inequity averse or reciprocal subjects in future periods. However, this incentive is absent in a stranger treatment where all interactions are one-shot so that there are no future gains. Note that this theory also explains that in a partner treatment cooperation declines over time but restarts again if subjects play another finitely repeated game[18]. And because the theories explain why people punish free-riders they can account for the punishment-related facts (6) – (8).

In summary, social preferences for fairness/equity, reciprocity or a prosocial self-image and the desire to avoid guilt are likely to play an important role in norm compliance. They provide an intrinsic motive to obey the normative standard to some extent and/or to sanction those who violate it. All of these theories are consistent with the notion that emotions are a key driver of the social preference although – with the exception of guilt aversion theory which explicitly models the emotion of guilt – they do not

explicitly incorporate emotions in the model. Yet, if an individual has a non-pecuniary desire to implement a fair outcome or to punish somebody who behaves unfairly, this desire is perfectly compatible with the idea that the emotional warm glow of achieving fairness or the indignation resulting from the violation of a fairness norm is the driver behind the social preference.

Although social preferences help in achieving norm compliance it is important to distinguish them conceptually from social norms which are defined as widely shared and approved normative standards. These standards are the essence of a social norm and they affect social preferences by defining what is considered as fair/equitable, kind or prosocial but they are conceptually nevertheless distinct. In particular, social preferences – for example the willingness to give up material resources to implement a fair outcome – are a property of the individual while social norms are a property of the group. For instance, in some groups the social norm could prescribe that the equal split is fair whereas in a more hierarchically organized group some degree of inequality may be required to satisfy the fairness norm. Technically, the fairness norm enters an individual's utility function as an exogenous parameter (at the individual level) that determines the individual's behavior together with the other parameters of the utility function. Once the parameters of an individual's utility function and the fairness norm are known and stable, this approach makes clear predictions over a wide range of situations. Of course, to the extent to which fairness norms vary across contexts, assuming a stable fairness norm causes wrong predictions. Nevertheless, for the sake of generating refutable predictions theories of fairness and reciprocity have typically taken a clear stance on the definition of fairness and kindness.

The direct norm approach on the other hand is silent about the underlying motives that induce individuals to comply with a prevailing social norm and theoretical papers that apply this approach[54] often make ad-hoc assumptions about the social norm while empirical studies do not define ex-ante the content of the normative standard but instead measure the norm empirically[55,73]. This renders the direct norm approach more flexible and more difficult to refute unless it is possible to reliably identify the normative standard empirically over the relevant range of situations.

How can we identify social norms?

There are several methods for the identification of social norms[24,55,74-76]. One method builds on the premise that humans are willing to incur personal costs to sanction the violation of a norm even if they are not directly hurt by the violation. One reason for this willingness may be that norm violations have been shown to cause indignation or even outrage[23,77,78] and these emotions may provide the raw material for the willingness to punish. Another reason may be that norm violators are typically perceived to deserve punishment[79] and, therefore, sanctioning them provides satisfaction – a hypothesis that is consistent with the finding that reward-related brain areas are activated during the

punishment of norm violators[80] and that already preschool children and chimpanzees are willing to pay for watching the punishment of antisocial actors[81].

Whatever, the precise reason may be, if norm violations trigger the desire to punish the perpetrators, we have a potential tool for identifying the norm as those behavior that is not punished by uninvolved third parties. Various studies have therefore employed a third party punishment paradigm for the study of social norms[24,82-85]. To identify a potential conditional cooperation norm[24], the third party had the option to punish the players in a two-person prisoners' dilemma game after she observed the actions of the two players. Every punishment point assigned to a player costs the third party one money unit (MU) and the sanctioned party three MUs. If the conditional cooperation norm applies to this game we should observe that unilateral defection is more deserving of punishment than bilateral defections and that bilateral cooperation is not at all or less punished than bilateral defection. This is exactly what the data shows[24]: 20.8% of third parties punish and the average expenditure on sanctioning is 0.6 MUs when both players defect but when a defector is paired with a cooperator the third party incurs much higher punishment cost of 3.4 MUs on average and the frequency with which they punish is also much higher (45.8%). In contrast, bilateral cooperation is almost never punished (4.2%) and the average punishment expenditure is negligible (0.08 Mus). These findings illustrate that the same action – free-riding – is perceived as normatively very different depending on whether the other player cooperated – a finding that is replicated in two recent studies[86,87]. Another, survey-based, study[74] examines subjects' moral judgments of complete defection in a two-person social dilemma as a function of the other person's cooperation level. Individuals' moral judgments show again the pattern one would expect if a norm of conditional cooperation applies. Defecting on a defecting player is perceived as much less or not at all immoral while the condemnation of a defection on a cooperative player strongly increases with that player's cooperation level.

An interesting method for the identification of social norms is based on the idea that social norms provide a focal point such that subjects' normative judgements are coordinated on this focal point[55]. This approach provides an incentivized measure of social norms by asking subjects to rate the extent to which an action is "socially appropriate and consistent with moral or proper social behavior," or "socially inappropriate and inconsistent with moral or proper social behavior." Subjects choose between "very socially inappropriate," "somewhat socially inappropriate," "somewhat socially appropriate" and "very socially appropriate." Subjects are not asked to provide their own personal evaluation, but to indicate what they believe is the most common answer, and they earn a monetary reward if their rating coincides with the modal answer of others. Thus, subjects have a monetary incentive to match other subjects' answers and if the social norm indeed constitutes a focal point the subjects have an incentive to assign the "very socially appropriate" rating – or at least the somewhat socially appropriate rating – to the behavior prescribed by the social norm. Thus, if a social norm exists

the distribution of appropriateness ratings has a peak at the normatively desired behavior. In contrast, if no social norm exists, the distribution of appropriateness ratings should be flat. This method has already been employed in several studies to elicit the social norm in social dilemmas[41,88] and to discriminate between the predictive power of the direct social norms approach and inequity aversion[88]. One of these studies[41] allows assessing the existence of a conditional cooperation norm in a public good game by directly eliciting subjects' view of the appropriateness of contributions conditional on other group members' behavior. Interestingly, this study shows that zero defection is the socially most appropriate behavior regardless of what other group members have contributed, suggesting that full cooperation has the largest normative appeal. However, if the average contribution of the other group members is low it becomes much more socially appropriate to reduce the cooperation rate which is consistent with the notion of conditional cooperation.

The focal point method of norm identification can also be used to answer the question whether the punishment of unfair behavior is a social norm. One study[89] applied the method to measure whether the punishment of unfair proposers in the ultimatum game – by rejecting their offer – is a social norm. Interestingly, the study shows that this is clearly not the case. Although nobody has yet examined whether the punishment of free-riders in a public goods game is a social norm the results of the ultimatum game study are consistent with the view that punishment of unfair behaviors – of which the sanctioning of free-riders is a special case – is not itself a social norm. Rather, the desire to punish free-riders seems to derive from other motivates such as to avoid inequity[63,90] or to reciprocate to unfair actions[68,87].

Another method for the identification of social norms in social dilemma games has recently been presented in two papers[91,92]. Here, each subject of the group is asked to indicate what other group member *should* contribute to the public good. The average of subjects' normative requests is afterwards conveyed to all group members and is likely to constitute a general standard of cooperation because it is commonly known and reflects the group members' views. Moreover, the higher subjects' agreement in their normative requests the more the average request will constitute a legitimate normative standard[92]. This approach enables the identification of the level of the prescribed behavior and the normative consensus among the group members which determines the legitimacy or strength of the norm. A strong social norm forms when group members give similar answers. Weaker social norms are characterized by some disagreement about the appropriate behavior. One advantage of this method is that it can be easily implemented in every period of a public goods game such that the level and the strength of the norm can be identified continuously. To what extent does this method support the existence of a conditional cooperation norm? The data show that the average *requested* contribution in a period is declining in subjects' average *actual* contributions in the previous period. Hence, this novel approach supports the evidence for the conditional cooperation norm found by the

other methods. In addition, the data show that when direct targeted punishment of free-riders is possible, subjects strongly obey the average normative request in their actual cooperation choices[92].

Thus, taken together, there is ample and diverse evidence for the existence of a conditional cooperation norm in social dilemma situations while there is little or no evidence that punishment of free-riders constitutes a social norm. These results show that one can provide discipline to the direct social norms approach and they strengthen the conjecture that a conditional cooperation norm shapes human cooperation. However, these norm elicitation approaches do not yet prove that cooperation behavior is *causally* affected by social norms because they – so far – only establish a correlation between the social norm and actual cooperative behavior[41].

Do social norms causally affect cooperation behavior?

The potential causal effect of social norms on behavior has been studied in various ways. A prominent approach[93,94] assumes that social norms need to be activated, i.e., become the focus of subjects' attention to affect behavior. Based on this view, a causal effect of social norms can be identified by varying the salience of the norm with various priming techniques. This literature shows that when subjects' attention is shifted towards social norms they begin to act in a more norm-congruent way[93-97]. For example, in one study[94] car drivers, who did not know that they were part of an experiment, saw the following handbill on their windshield when they returned to their car located on a large parking lot: "April is Keep Arizona Beautiful Month. Please Do Not Litter". In a second condition, the text on the handbill was "April is Conserve Arizona's Energy Month. Please Turn Off Unnecessary Lights", and in a third (control) condition they could read "April is Arizona's Fine Arts Month. Please Visit Your Local Art Museum". These treatments are likely to prime a relatively strong anti-littering norm in the first case, a weaker anti-littering norm in the second case and no norm in the third case. In line with the hypothesis that a stronger activation of the anti-littering norm leads to less littering, car drivers threw the handbill on the ground in only in 10% of the cases in the first treatment, in 18% of the cases in the second condition and in 25% of the cases in the third condition. Findings like these raise the question which aspect of the social norm is the causal driver of the behavior change. Does the increase in the salience of the norm change the social appropriateness rating of norm compliant behavior? Or does it merely change subjects' views about how widely the norm is shared? Or does it change subjects' feelings of guilt if they litter? Unfortunately, we do not know the answer to these questions.

Another literature studies the effects of communication on cooperation in settings where subjects have the chance to communicate before the play of a social dilemma game. Face-to-face communication increases cooperation and there are potentially many channels through which this can occur[8,10,11,98] but one reason could be that subjects often make cooperation promises in their pre-play

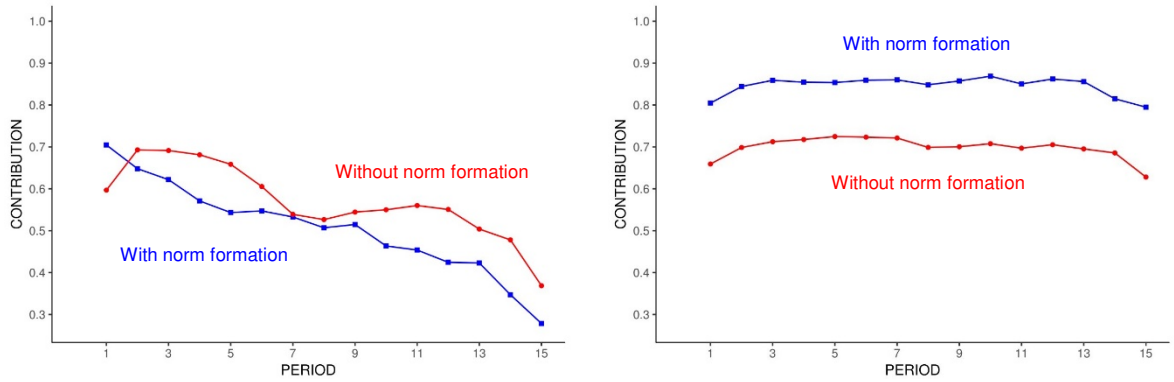
interactions. Therefore, if a social norm of keeping promises is at work communication increases cooperation[56]. One study[99] explicitly examined the role of promises in communication by allowing subjects' to communicate via computer the numerical contribution level in an entry labelled "possible contributions". Other group members observed these numerical messages and all group members could interactively alter their entries within a certain time frame. This type of numerical communication generally had no effect on cooperation unless two further features were simultaneously present. First, subjects had to have the possibility to impose monetary punishment on others and second, they needed to have the possibility to promise to contribute their "possible contribution" entry. Unfortunately, none of these communication studies directly identified the normatively desired contribution level, which limits the scope for the establishment of a direct link between communication and social norms.

A recent study[73] solves this problem by eliciting the socially most appropriate action in two person social dilemmas in the presence and in the absence of an informal agreement to choose the cooperation level that maximizes the joint payoff. No agreements on other cooperation levels were possible. The results show that informal, yet unenforceable, agreements to choose the joint payoff maximizing action strongly increases the social appropriateness of that action relative to all other actions and causes more cooperative behavior. There is thus a clear empirical link between the informal agreement, i.e., the promise to choose the joint payoff-maximizing action, the increase in the social norm and the increase in cooperation rates. Although this study clearly shows that the standard of behavior (social appropriateness) is causally affected by informal promises, it does not show that this change in the standard of what is appropriate directly causes the change in behavior. The reason is that the change in the standard could also have opened other psychological channels of behavior change. For example, it could be the case that the increase in the normative standard causes an increase in what subjects believe that their partner will expect from them and this change in the expectation could induce guilt and thus norm compliance could be triggered via guilt aversion.

The above-mentioned method for norm identification[91,92], that relies on subjects' period-by-period normative requests, can also be used to study the causal impact of social norms on behavior by comparing treatments in which subjects have the opportunity to announce normative requests with treatments where this opportunity is absent. In treatments with normative requests the average request constitutes a commonly known standard of behavior that is absent in treatments without normative requests. In one study[92] the authors introduce the norm formation opportunity in finitely repeated public goods games with partner matching where the possibility to punish other group members is either absent or present. Interestingly, when the possibility of punishment is absent, the opportunity to form a normative standard has no impact on behavior while in the presence of the

possibility to punish the normative standard causes a significant and stable increase in cooperation rates (Figure 2).

This radically different impact of social norms on cooperation conditional on the existence of punishment opportunities exists despite the fact that the normative standard in the punishment and no-punishment treatment is very high and statistically indistinguishable during the first three periods. Nevertheless, there occur substantial norm deviations in the absence of punishment from the very beginning while in the presence of punishment the norm is largely obeyed throughout the whole experiment. Thus, the existence of a normative standard that renders high cooperation the socially most appropriate action, and focusses attention on the normative standard, is per se not sufficient to induce a change in cooperation behavior, suggesting that intrinsic motives for norm compliance are not sufficiently strong and that the punishment threat is needed to establish a stable norm-driven behavior change in a population of heterogeneously motivated actors.



(a) Public goods game without peer punishment (b) Public goods game with peer punishment

Figure 2 The effect of social norms with and without punishment

Notes: This graph is taken from Fehr and Schurtenberger [92]. Fixed groups of four subjects play a public good game over 15 periods. The graph shows average normalized contributions over time, that is, 1 corresponds to full contributions and 0 to no contributions. Treatments with a punishment opportunity also allow for the counter-punishment of those who punish free-riders to examine whether norms have a causal impact in an environment that has been shown to be hostile for human cooperation[32].

Normative constraints and the (in)efficiency of peer punishment

The existence of punishment opportunities in public goods games causes strong cooperation increases in Western but not in all cultures[31,100,101]. In particular, in those countries that have weak norms of civic cooperation – defined as the willingness to evade taxes, make fraudulent claims to receive welfare state benefits or dodging fares on public transport – the antisocial punishment of cooperators is particularly strong and is associated with detrimental effects on overall cooperation rates. This finding is consistent with the view that norms of civic cooperation have a causal, constraining, effect on

antisocial punishment. However, the finding does not prove causality because there could be other reasons that may account for the correlation between antisocial punishment and norms of civic cooperation. For example, countries with low norms of civic cooperation often also have bad schools (e.g., because of teacher absenteeism or low teacher quality[102,103]) and school or teacher quality might shape both norms of civic cooperation and restraints on antisocial punishment.

Although the antisocial punishment of above-average cooperators by those who cooperate less tends to be rare in Western cultures, it has been observed from the beginning and several potential reasons for its existence have been mentioned[22]. First, in rare cases, it may simply reflect a random choice error. Second, there is evidence that a small yet significant proportion of subjects regularly displays envious or spiteful motives[104,105], implying that they prefer to spend money to hurt others regardless of their level of prosociality. Third, antisocial punishment may be the result of a coordination failure among reciprocally motivated subjects that are in principle willing to cooperate. Consider a reciprocal subject with pessimistic beliefs about others' cooperation. These subjects may cautiously start with an intermediate or low level of cooperation while other subjects have optimistic expectations, start with high cooperation and punish those who cooperate less. The pessimistic, yet willing, low contributor may view this as an unfair punishment and may thus retaliate in the next period against the high contributors. These events may spoil the whole group and lead to a process of punishment and counter-punishment with detrimental effects on cooperation. In fact, if subjects are given explicit counter-punishment opportunities[30,32] some subjects use them to the detriment of the group's cooperation and welfare by punishing those who punished them for free-riding. More generally, public goods experiments that allow for peer punishment often fail to increase the overall welfare of the group members for an extended period of time despite the large increase in cooperation rates[22,25,38]. The reason for this is the high collateral cost associated with peer punishment.

However, the very fact that peer punishment can get out of control suggests that societies have developed mechanisms to constrain and control it. After all, peer punishment is physically always possible when two or more individuals directly interact with each other. It appears impossible for society to ever control or constrain all the different forms of peer punishment – that range from a raised eye brow or verbal insult to mobbing, ostracism, public shaming and corporal punishment – except through the normative control of people's behavior. Interestingly, the literature on simple societies[106,107] provides ample evidence on the ways in which societies impose constraints on punishment. One study[107], for example, reports how the Ju/'hoansi bushmen, a group of hunter-gatherers living in Botswana subject peer punishment to strong habitual and normative constraints: if a man is publicly criticized for norm violations this is often done by a women to avoid the escalation of arguments among men.

In view of the normative constraints that societies impose on the permissible forms of punishment it is interesting to study whether subjects voluntarily impose on themselves institutions that regulate punishment by either ruling out peer punishment completely[108], replace it by a centralized state that automatically imposes taxes to finance public goods[109] or by an enforcement mechanism that rules out antisocial peer punishment[110-113]. If given the chance, subjects often prefer institutions that rule out antisocial peer punishment but how, in practice, is it possible to achieve this without also ruling out peer punishment altogether. More fundamentally, how is it ever possible to rule out peer punishment altogether in a world in which people socially interact with each other and in which the centralized legal enforcement of rules is always imperfect. Thus, because peer punishment is always possible to some degree, a key question is how we can constrain it *through normative standards*.

This question can be answered by comparing the punishment patterns in settings with and without the opportunity for normative requests[92]. It turns out that when subjects can form a normative cooperation standard the punishment of free-riders becomes less severe. Thus, the normative standard increases cooperation while simultaneously decreasing the punishment of free-riders, suggesting that the punishment of free-riders becomes more effective. In fact, punished free-riders indeed increase their cooperation subsequently more strongly when the normative standard is present[92]. In addition, antisocial punishment also decreases in the presence of a normative cooperation standard thus lending support to the hypothesis that norms of civic cooperation may causally reduce antisocial punishment.

These results are obtained despite the existence of counter-punishment opportunities which have previously – in the absence of normative requests – led to very low levels of cooperation and welfare. If subjects have the chance to coordinate on a normative consensus by establishing a cooperation standard the existence of peer punishment opportunities is from the very beginning associated with a higher group welfare compared to a setting in which peer punishment is ruled out[92].

In view of the high collateral cost of normatively uncoordinated peer punishment it is remarkable that in the presence of migration opportunities basically all subjects eventually migrate from a setting without any targeted punishment opportunities to one with uncoordinated peer punishment (fact 8). However, in these experiments the only alternative to uncoordinated peer punishment was the complete absence of punishment opportunities. What happens if we additionally allow migration to the following two institutions: (i) normatively coordinated peer punishment and (ii) normative coordination and punishment by a central (democratically elected) authority?

When given the chance to migrate between all four institutions – no punishment, uncoordinated peer punishment, normatively coordinated peer punishment and centralized punishment with normative coordination – subjects basically *never* enter the uncoordinated peer punishment institution and

rather quickly almost all subjects enter institutions with normative coordination (Fig. 3). These institutions minimize or fully eradicate antisocial punishment and generate high levels of cooperation without the collateral damages associated with uncoordinated peer punishment[91]. This demonstrates that the traditional uncoordinated peer punishment institution fails to capture a very important dimension: the strong demand for normative coordination and regulation – a demand that societies who inevitably have to rely on some forms of peer sanctioning typically satisfy through the formation of social norms that put constraints on individuals' sanctioning behavior. Of course, groups will not automatically solve inefficient peer sanctioning through informal constraints but it seems likely that those groups who do solve this problem in a more efficient way will be more successful because they are better able to solve their collective action problems[1,114,115]. Therefore, they are better able to compete with other groups. Thus, conclusions regarding the effectiveness and the welfare properties of peer punishment may provide a misleading picture if they are based on institutional settings that rule out suitable normative consensus building opportunities that can put constraints on peer sanctioning.

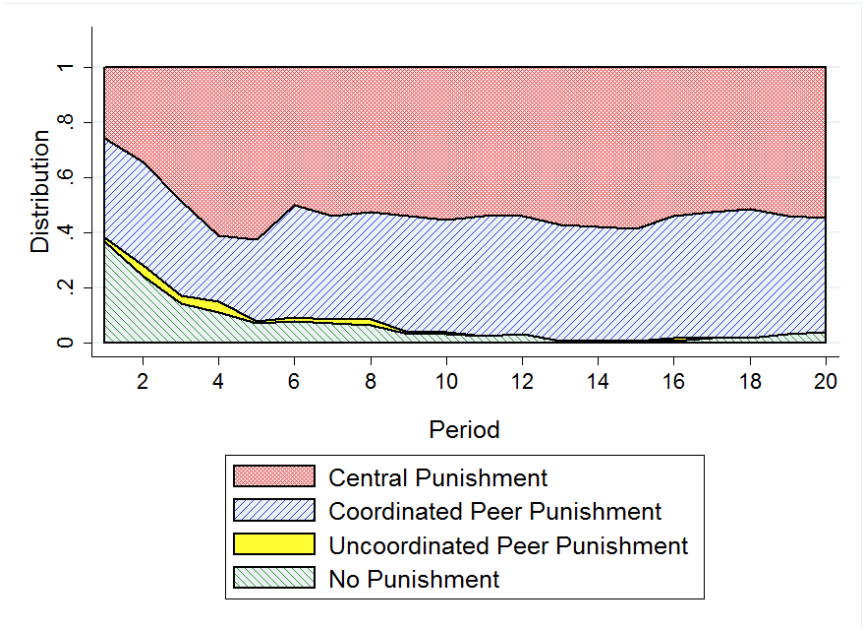


Figure 3: Norm formation and the emergence of efficient punishment institutions

Notes: This graph is taken from Fehr and Williams [91]. The figure depicts how the share of subjects in the four available institutions evolves over time. "No Punishment" is a regular public goods game without punishment. "Uncoordinated Peer Punishment" is a public goods game with peer punishment, but without norm formation. "Coordinated Peer Punishment" is a public goods game with the possibility of peer punishment and a norm formation opportunity. "Central Punishment" is a public goods game with norm formation and punishment exerted by an elected authority.

Summary and open questions

The pervasiveness of social norms and the ubiquity of cooperation among non-kin are two salient features of human societies. Many social norms are beneficial for overall society and to the extent to which individuals voluntarily obey them we can view their behavior as an act of cooperation. Although humans are by no means the only species displaying cooperation among individuals, it has often been pointed out that the breadth and depth of human large-scale cooperation among non-kin in a globalized world, as well as the observed cooperation in one-shot encounters, appear unique in the animal kingdom[5,116-118]. Several potential factors – such as limited memory or excessive time discounting[117,119] – may constitute evolutionary obstacles to cooperation in animal species but perhaps the cognitive prerequisites for social norms are also relevant. For example, the very notion of a normative standard – what ought to be done – is rather complex and perhaps even impossible to identify reliably in species that lack sophisticated language. The same applies to the notion of normative approval and disapproval. Therefore, it is perhaps not surprising that our closest living relatives do not seem to share some of our most fundamental norms of fairness and cooperation[120-122] (although see [123,124]) and that there seems to be no evidence for third party punishment of norm violations harming non-kin in non-human species[125]. In contrast, third party punishment of non-kin and even strangers is frequent in humans[24,59,82] and young children already have a working knowledge of social norms[122,126,127]. The widespread prevalence of social norms may therefore well be one of the defining characteristics of our species and a crucial determinant of human cooperation.

The evidence suggests that human cooperation is strongly affected by normative considerations. Various methods for the identification of social norms indicate the existence of a conditional cooperation norm that is a key ingredient for the explanation of major cooperation-related facts. The behavioral strength of the conditional cooperation norm probably also derives from the fact that it is consistent with principles of equity and reciprocity and that purely conformist preferences also imply conditionally cooperative behavior. Compliance with social norms relies on the existence of social preferences that incorporate abstract normative principles such as equity or reciprocity – which also provide foundations for the willingness to punish norm violators – or are based on the desire for avoiding disapproval, a prosocial self-image or the avoidance of disappointing others. Social norms also appear to guide and constrain punishment behavior and subjects have a strong desire for environments that enable normative coordination.

There are, however, still many important unanswered questions. Reliable empirical knowledge about the precise channels through which norms have a causal impact is, for example, still scarce. Does the normative standard shape behavior directly via an intrinsic utility component or does it have an impact by affecting and coordinating beliefs about others' cooperation. Or does it guide the punishment of

free-riders and affect beliefs about punishment in case of non-compliance? In addition, there are many other intriguing and exciting questions that are awaiting an answer (see text box on important unresolved research problems), implying that there is still much to discover in this area of research.

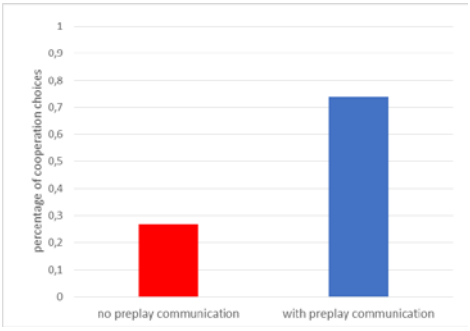


Fig. 1a. Cooperation rates in a one-shot social dilemma game with and without pre-play communication among the subjects[8]. This graph illustrates Result (1).

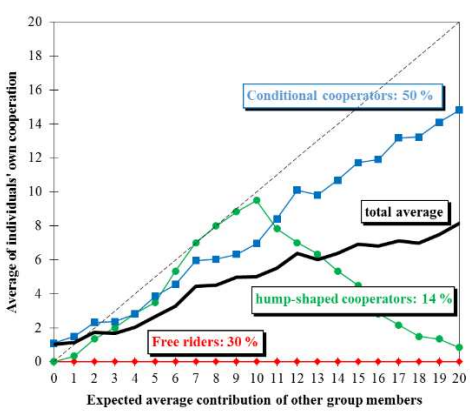


Fig. 1b. Evidence for Result 2. Higher expectations of other group members' cooperation causes on average an increase in individual's own cooperation but individuals are heterogeneous with 50% conditional cooperators, 30% full free-riders and 14% hump-shaped conditional cooperators[12].

Fig. 1c: Evidence for Result (3). Decline in cooperation rates over time in finitely repeated public goods experiments in which free-riding is the payoff maximizing strategy for selfish subjects[22,128,129].

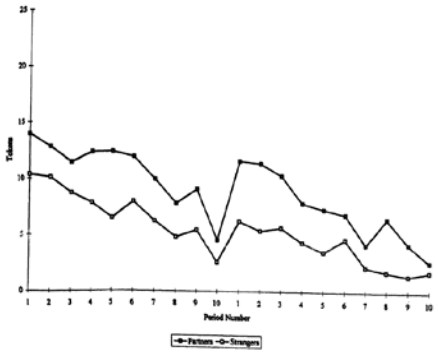


Fig. 1d. Evidence for Result (3) and (4). Cooperation rates in partner treatments are typically higher than those in stranger treatments. In this study, subjects initially believed to interact for 10 periods after which the experimenter implemented a surprise restart of the same 10-period experiment.[19]

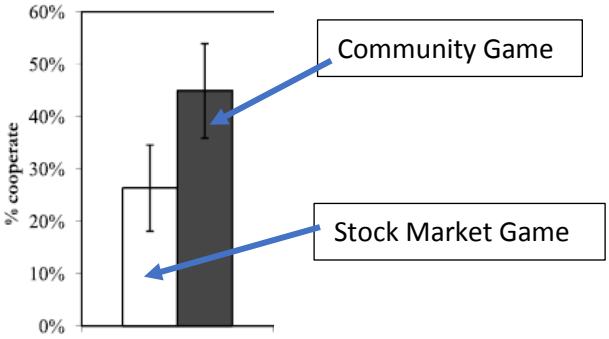


Fig. 1e. Evidence for Result (5). Merely calling the prisoners' dilemma a Community game – as opposed to a Stock Market game – increases cooperation; but if the game is played sequentially, this framing effect vanishes[20].

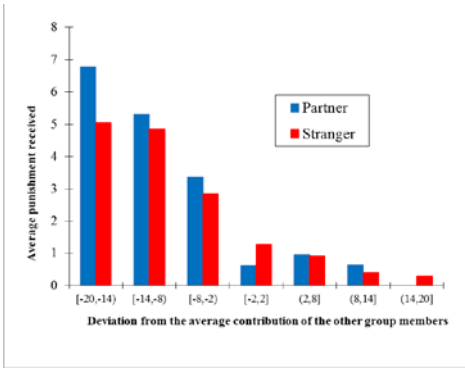


Fig. 1f. Evidence for Result (6) and (7). Punishment – measured in terms of the experienced percentage reduction in income – of group members as a function of the deviation of their cooperation level from the average cooperation of other group members. Punishment of free-riders is very high – even in the stranger treatment – but above average cooperators also face some "perverse" punishment[22].

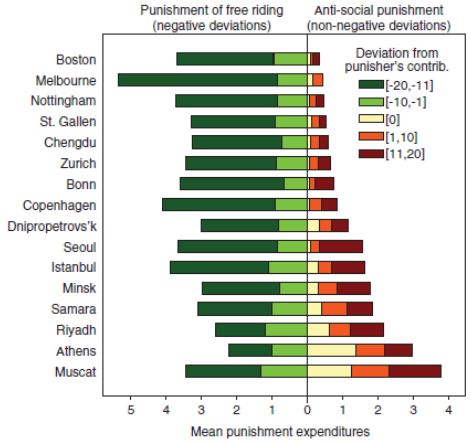


Fig. 1g Evidence for Result (7). There are strong cultural differences in antisocial punishment of cooperators[31].

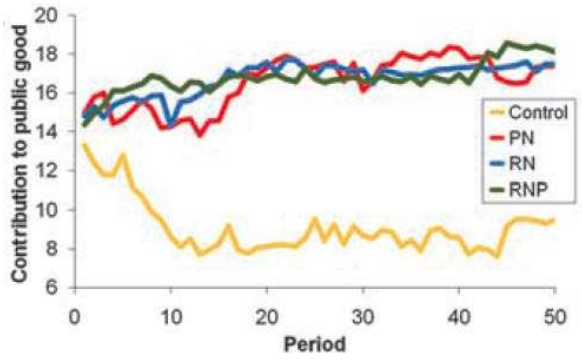


Fig 1i. Evidence for Result (6) and (9). The opportunity to punish peers after they observed others' cooperation levels (treatment PN) leads to large increases in cooperation relative to a control treatment without peer punishment (Control). The opportunity to mutually reward each other (RN) leads to similarly high cooperation levels compared to PN and treatments with both reward and punishment (RNP)[38].

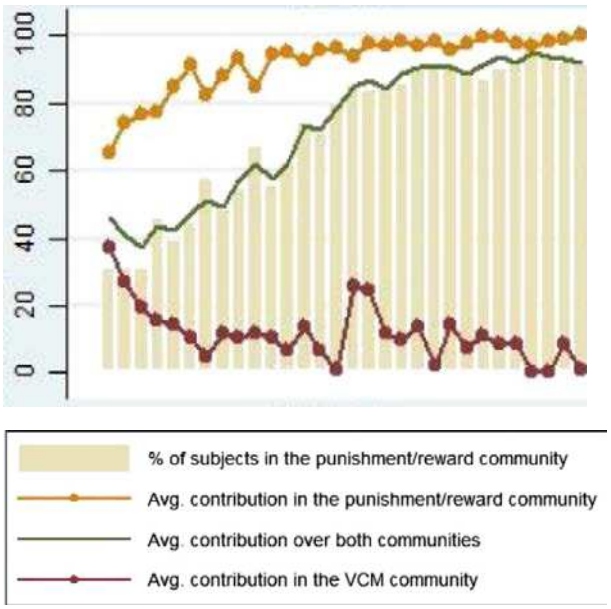


Fig. 1h. Evidence for Result (8). Subjects could choose in every period whether they want to be in the group with a peer punishment opportunity or the group without this opportunity (VCM). The figure shows that the vast majority of subjects eventually prefer the community with peer punishment[34].

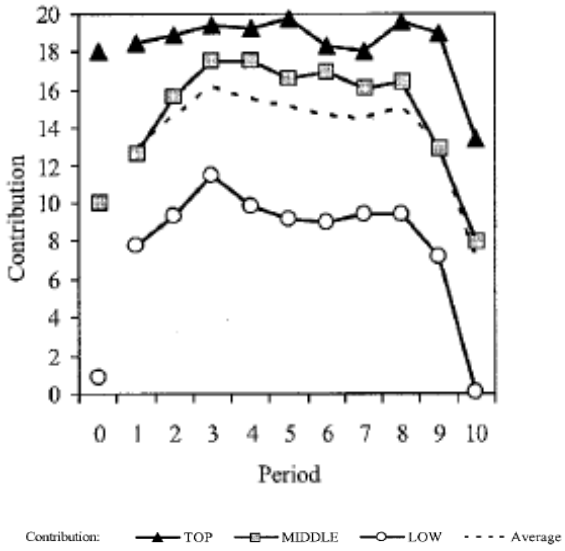


Fig. 1j. Evidence for Result (10i). Top cooperators in a one-shot PD are grouped together in a subsequent 10-period public goods game. Likewise, the middle and the low cooperators are grouped together. Top cooperators achieve very high cooperation rates during the first nine periods[40].

1 Boyd, R. & Richerson, P. J. The Evolution of Norms - an Anthropological View. *Journal of Institutional and Theoretical Economics* **150**, 72-87, (1994).
2 Sober, E. & Wilson, D. S. *Unto others: The evolution and psychology of unselfish behavior*. (Harvard University Press, 1999).
3 Elster, J. *The cement of society: A survey of social order*. (Cambridge University Press, 1989).
4 Hobbes, T. *Leviathan*. (Continuum, 2005 Orig. pub. 1651).
5 Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785, (2003).
6 Fehr, E. & Fischbacher, U. Social norms and human cooperation. *Trends in cognitive sciences* **8**, 185-190, (2004).
7 Bicchieri, C. *The grammar of society: The nature and dynamics of social norms*. (Cambridge University Press, 2006).
8 Dawes, R. M., McTavish, J. & Shaklee, H. Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of personality and social psychology* **35**, 1, (1977).
9 Dawes, R. M. Social dilemmas. *Annual review of psychology* **31**, 169-193, (1980).
10 Isaac, R. M. & Walker, J. M. Communication and free-riding behavior: The voluntary contribution mechanism. *Econ Inq* **26**, 585-608, (1988).
11 Sally, D. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and society* **7**, 58-92, (1995).
12 Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ Lett* **71**, 397-404, (2001).
13 Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J. & Sutter, M. Conditional cooperation on three continents. *Econ Lett* **101**, 175-178, (2008).
14 Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* **14**, 47-83, (2011).
15 Isaac, M. R., McCue, K. & Plott, C. R. Public Goods Provision in an Experimental Environment. *Journal of Public Economics* **26**, 51-74, (1985).
16 Kim, O. & Walker, J. M. The Free Rider Problem: Experimental Evidence. *Public Choice* **43**, 3-24, (1984).
17 Andreoni, J. Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics* **110**, 1-21, (1995).
18 Ambrus, A. & Pathak, P. A. Cooperation over finite horizons: A theory and experiments. *Journal of Public Economics* **95**, 500-512, (2011).
19 Croson, R. Partners and Strangers Revisited. *Econ Lett* **53**, 25-32, (1996).
20 Ellingsen, T., Johannesson, M., Mollerstrom, J. & Munkhammar, S. Social framing effects: Preferences or beliefs? *Game Econ Behav* **76**, 117-130, (2012).
21 Liberman, V., Samuels, S. M. & Ross, L. The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin* **30**, 1175-1185, (2004).
22 Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980-994, (2000).
23 Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137-140, (2002).
24 Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol Hum Behav* **25**, 63-87, (2004).
25 Gächter, S., Renner, E. & Sefton, M. The long-run benefits of punishment. *Science* **322**, 1510-1510, (2008).
26 Fehr, E. & Rockenbach, B. Detrimental effects of sanctions on human altruism. *Nature* **422**, 137-140, (2003).
27 Houser, D., Xiao, E., McCabe, K. & Smith, V. When punishment fails: Research on sanctions, intentions and non-cooperation. *Game Econ Behav* **62**, 509-532, (2008).
28 Xiao, E. T. Profit-seeking punishment corrupts norm obedience. *Game Econ Behav* **77**, 321-344, (2013).

29 Fehr, E. & List, J. A. The Hidden Costs and Returns of Incentives-Trust and Trustworthiness among Ceos. *J Eur Econ Assoc* **2**, (2004).
30 Cinyabuguma, M., Page, T. & Putterman, L. Can second-order punishment deter perverse punishment? *Experimental Economics* **9**, 265-279, (2006).
31 Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362-1367, (2008).
32 Nikiforakis, N. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* **92**, 91-112, (2008).
33 Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108-111, (2006).
34 Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. On cooperation in open communities. *Journal of Public Economics* **120**, 220-230, (2014).
35 Brown, M., Falk, A. & Fehr, E. Relational contracts and the nature of market interactions. *Econometrica* **72**, 747-780, (2004).
36 Rockenbach, B. & Milinski, M. The efficient interaction of indirect reciprocity and costly punishment. *Nature* **444**, 718-723, (2006).
37 Sefton, M., Shupp, R. & Walker, J. M. The effect of rewards and sanctions in provision of public goods. *Econ Inq* **45**, 671-690, (2007).
38 Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D. & Nowak, M. A. Positive interactions promote public cooperation. *Science* **325**, 1272-1275, (2009).
39 Balliet, D., Mulder, L. B. & Van Lange, P. A. M. Reward, Punishment, and Cooperation: A Meta-Analysis. *Psychological Bulletin* **137**, 594-615, (2011).
40 Gächter, S. & Thöni, C. Social learning and voluntary cooperation among like-minded people. *J Eur Econ Assoc* **3**, 303-314, (2005).
41 Kimbrough, E. O. & Vostroknutov, A. Norms make preferences social. *J Eur Econ Assoc* **14**, 608-638, (2016).
42 Chaudhuri, A., Graziano, S. & Maitra, P. Social learning and norms in a public goods experiment with inter-generational advice. *Review of Economic Studies* **73**, 357-380, (2006).
43 Barr, A., Packard, T. & Serra, D. Participatory accountability and collective action: Experimental evidence from Albania. *European Economic Review* **68**, 250-269, (2014).
44 Fehr, E. & Leibbrandt, A. A field study on cooperativeness and impatience in the tragedy of the commons. *Journal of Public Economics* **95**, 1144-1155, (2011).
45 Rustagi, D., Engel, S. & Kosfeld, M. Conditional cooperation and costly monitoring explain success in forest commons management. *Science* **330**, 961-965, (2010).
46 Keizer, K., Lindenberg, S. & Steg, L. The Spreading of Disorder. *Science* **322**, 1681-1685, (2008).
47 Kosfeld, M. & Rustagi, D. Leader punishment and cooperation in groups: Experimental field evidence from commons management in Ethiopia. *American Economic Review* **105**, 747-783, (2015).
48 Breza, E., Kaur, S. & Krishnaswamy, N. Scabs: norm-driven suppression of labor supply. *working paper*, (2018).
49 Kaur, S. Nominal Wage Rigidity in Village Labor Markets. *American Economic Review*, (forthcoming).
50 Gelcich, S., Guzman, R., Rodríguez-Sickert, C., Castilla, J. C. & Cárdenas, J. C. Exploring external validity of common pool resource experiments: insights from artisanal benthic fisheries in Chile. *Ecology and Society* **18**, (2013).
51 Burks, S. *et al.* Lab measures of other-regarding preferences can predict some related on-the-job behavior: Evidence from a large scale field experiment. *working paper*, (2016).
52 Carlsson, F., Johansson-Stenman, O. & Nam, P. K. Social preferences are stable over long periods of time. *Journal of Public Economics* **117**, 104-114, (2014).
53 Ostrom, E. Collective action and the evolution of social norms. *J Econ Perspectives* **14**, 137-158, (2000).
54 Lindbeck, A., Nyberg, S. & Weibull, J. W. Social norms and economic incentives in the welfare state. *Quarterly Journal of Economics* **114**, 1-35, (1999).
55 Krupka, E. L. & Weber, R. A. Identifying social norms using coordination games: Why does dictator game sharing vary? *J Eur Econ Assoc* **11**, 495-524, (2013).

56 Bicchieri, C. Covenants without swords: Group identity, norms, and communication in social dilemmas. *Rationality and Society* **14**, 192-228, (2002).
57 Fischbacher, U. & Gächter, S. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review* **100**, 541-556, (2010).
58 Barclay, P. Reputational benefits for altruistic punishment. *Evol Hum Behav* **27**, 325-344, (2006).
59 Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473-476, (2016).
60 Kiyonari, T. & Barclay, P. Cooperation in social dilemmas: Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology* **95**, 826-842, (2008).
61 Rockenbach, B. & Milinski, M. To qualify as a social partner, humans hide severe punishment, although their observed cooperativeness is decisive. *P Natl Acad Sci USA* **108**, 18307-18312, (2011).
62 Wrong, D. H. The Oversocialized Conception of Man in Modern Sociology. *American Sociological Review* **26**, 183-193, (1961).
63 Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* **114**, 817-868, (1999).
64 Bolton, G. E. & Ockenfels, A. ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, 166-193, (2000).
65 Lopez-Perez, R. Aversion to norm-breaking: A model. *Game Econ Behav* **64**, 237-267, (2008).
66 Rabin, M. Incorporating fairness into game theory and economics. *American Economic Review*, 1281-1302, (1993).
67 Dufwenberg, M. & Kirchsteiger, G. A theory of sequential reciprocity. *Game Econ Behav* **47**, 268-298, (2004).
68 Falk, A. & Fischbacher, U. A theory of reciprocity. *Game Econ Behav* **54**, 293-315, (2006).
69 Battigalli, P. & Dufwenberg, M. Guilt in games. *American Economic Review* **97**, 170-176, (2007).
70 Dufwenberg, M., Gachter, S. & Hennig-Schmidt, H. The framing of games and the psychology of play. *Game Econ Behav* **73**, 459-478, (2011).
71 Dhami, S., Wei, M. & Al-Nowaihi, A. Public goods games and psychological utility: theory and evidence. *J Econ Behav Organ*, (forthcoming).
72 Benabou, R. & Tirole, J. Identity, Morals, and Taboos: Beliefs as Assets. *Quarterly Journal of Economics* **126**, 805-855, (2011).
73 Krupka, E. L., Leider, S. & Jiang, M. A meeting of the minds: informal agreements and social norms. *Management Science*, (2016).
74 Cubitt, R. P., Drouvelis, M., Gächter, S. & Kabalin, R. Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics* **95**, 253-264, (2011).
75 Reuben, E. & Riedl, A. Enforcement of contribution norms in public good games with heterogeneous populations. *Game Econ Behav* **77**, 122-137, (2013).
76 Bicchieri, C. *Norms in the Wild*. (Oxford University Press, 2017).
77 Xiao, E. & Houser, D. Emotion expression in human punishment behavior. *P Natl Acad Sci USA* **102**, 7398-7401, (2005).
78 Bosman, R., Sutter, M. & van Winden, F. The impact of real effort and emotions in the power-to-take game. *J Econ Psychol* **26**, 407-429, (2005).
79 Carlsmith, K. M., Darley, J. M. & Robinson, P. H. Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of personality and social psychology* **83**, 284, (2002).
80 DeQuervain, D. *et al.* The neural basis of altruistic punishment. *Science* **305**, 1254-1258, (2004).
81 Mendes, N., Steinbeis, N., Bueno-Guerra, N., Call, J. & Singer, T. Preschool children and chimpanzees incur costs to watch punishment of antisocial others. *Nature Human Behaviour* **2**, 45-51, (2018).
82 Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767-1770, (2006).

83 Marlowe, F. W. *et al.* More 'altruistic' punishment in larger societies. *Proceedings of the Royal Society of London B: Biological Sciences* **275**, 587-592, (2008).
84 Lewisch, P. G., Ottone, S. & Ponzano, F. Free-riding on altruistic punishment? An experimental comparison of third-party punishment in a stand-alone and in an in-group environment. *Review of Law & Economics* **7**, 161-190, (2011).
85 Lergetporer, P., Angerer, S., Glätzle-Rützler, D. & Sutter, M. Third-party punishment increases cooperation in children through (misaligned) expectations and conditional cooperation. *Proceedings of the National Academy of Sciences* **111**, 6916-6921, (2014).
86 Kamei, K. Altruistic Norm Enforcement and Decision-Making Format in a Dilemma: Experimental Evidence. (2017).
87 Carpenter, J. P. & Matthews, P. H. Norm enforcement: anger, indignation, or reciprocity? *J Eur Econ Assoc* **10**, 555-572, (2012).
88 Gächter, S., Nosenzo, D. & Sefton, M. Peer effects in pro-social behavior: Social norms or social preferences? *J Eur Econ Assoc* **11**, 548-573, (2013).
89 Bartling, B. & Özdemir, Y. The Limits to Moral Erosion in Markets: Social Norms and the Replacement Excuse. *working paper*, (2017).
90 Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794-796, (2007).
91 Fehr, E. & Williams, T. Creating an Efficient Culture of Cooperation. *working paper*, (2017).
92 Fehr, E. & Schurtenberger, I. The Dynamics of Norm Formation and Norm Decay. *working paper*, (2017).
93 Cialdini, R. B., Kallgren, C. A. & Reno, R. R. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in experimental social psychology* **24**, 201-234, (1991).
94 Kallgren, C. A., Reno, R. R. & Cialdini, R. B. A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and social psychology bulletin* **26**, 1002-1012, (2000).
95 Berkowitz, L. & Daniels, L. R. Affecting the salience of the social responsibility norm: effects of past help on the response to dependency relationships. *The Journal of Abnormal and Social Psychology* **68**, 275, (1964).
96 Berkowitz, L. Social norms, feelings, and other factors affecting helping and altruism. *Advances in experimental social psychology* **6**, 63-108, (1972).
97 Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics* **148**, 14-31, (2017).
98 Ostrom, E., Walker, J. & Gardner, R. Covenants with and without a sword: Self-governance is possible. *American political science Review* **86**, 404-417, (1992).
99 Bochet, O. & Putterman, L. Not just babble: Opening the black box of communication in a voluntary contribution experiment. *European Economic Review* **53**, 309-326, (2009).
100 Gächter, S. & Herrmann, B. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **364**, 791-806, (2009).
101 Gächter, S. & Herrmann, B. The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review* **55**, 193-210, (2011).
102 Hanushek, E. A. & Woessmann, L. Knowledge capital, growth, and the East Asian miracle Access to schools achieves only so much if quality is poor. *Science* **351**, 344-345, (2016).
103 Hanushek, E. A. & Rivkin, S. G. The Distribution of Teacher Quality and Implications for Policy. *Annual Review of Economics, Vol 4* **4**, 131-158, (2012).
104 Fehr, E., Hoff, K. & Kshetramade, M. Spite and development. *American Economic Review* **98**, 494-499, (2008).
105 Bruhin, A., Fehr, E. & Schunk, D. The Many Faces of Human Prosociality. *J Eur Econ Assoc*, (forthcoming).
106 Mathew, S. & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *P Natl Acad Sci USA* **108**, 11375-11380, (2011).

107 Wiessner, P. Norm enforcement among the Ju/'hoansi Bushmen - A case of strong reciprocity? *Human Nature-an Interdisciplinary Biosocial Perspective* **16**, 115-145, (2005).
108 Sutter, M., Haigner, S. & Kocher, M. G. Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies* **77**, 1540-1566, (2010).
109 Markussen, T., Putterman, L. & Tyran, J. R. Self-Organization for Collective Action: An Experimental Study of Voting on Sanction Regimes. *Review of Economic Studies* **81**, 301-324, (2014).
110 Yamagishi, T. The provision of a sanctioning system as a public good. *Journal of Personality and social Psychology* **51**, 110, (1986).
111 Ertan, A., Page, T. & Putterman, L. Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review* **53**, 495-511, (2009).
112 Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society of London B: Biological Sciences*, (2012).
113 Andreoni, J. & Gee, L. K. Gun for hire: delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics* **96**, 1036-1046, (2012).
114 Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and sociobiology* **13**, 171-195, (1992).
115 Henrich, J. Cultural group selection, coevolutionary processes and large-scale cooperation. *J Econ Behav Organ* **53**, 3-35, (2004).
116 Hammerstein, P. *Genetic and cultural evolution of cooperation*. (MIT press, 2003).
117 Stevens, J. R. & Hauser, M. D. Why be nice? Psychological constraints on the evolution of cooperation. *Trends in cognitive sciences* **8**, 60-65, (2004).
118 Boyd, R. & Richerson, P. in *Evolution and Culture* eds S. Levinson & P. Jaisson) (MIT Press, 2005).
119 Stephens, D. W., McLinn, C. M. & Stevens, J. R. Discounting and reciprocity in an iterated prisoner's dilemma. *Science* **298**, 2216-2218, (2002).
120 Jensen, K., Call, J. & Tomasello, M. Chimpanzees are rational maximizers in an ultimatum game. *science* **318**, 107-109, (2007).
121 Jensen, K., Call, J. & Tomasello, M. Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences* **104**, 13046-13050, (2007).
122 Ulber, J., Hamann, K. & Tomasello, M. Young children, but not chimpanzees, are averse to disadvantageous and advantageous inequities. *Journal of experimental child psychology* **155**, 48-66, (2017).
123 Proctor, D., Williamson, R. A., Waal, F. B. M. & Brosnan, S. F. Chimpanzees play the ultimatum game. *Proceedings of the National Academy of Sciences* **110**, 2070-2075, (2013).
124 Brosnan, S. F., Schiff, H. C. & De Waal, F. B. Tolerance for inequity may increase with social closeness in chimpanzees. *Proceedings of the Royal Society of London B: Biological Sciences* **272**, 253-258, (2005).
125 Riedl, K., Jensen, K., Call, J. & Tomasello, M. No third-party punishment in chimpanzees. *Proceedings of the National Academy of Sciences* **109**, 14824-14829, (2012).
126 McAuliffe, K., Jordan, J. J. & Warneken, F. Costly third-party punishment in young children. *Cognition* **134**, 1-10, (2015).
127 Cummins, D. D. Evidence of deontic reasoning in 3-and 4-year-old children. *Memory & Cognition* **24**, 823-829, (1996).
128 Isaac, M. R. & Walker, J. M. Divergent Evidence on Free Riding: An Experimental Examination of Some Possible Explanations. *Public Choice* **43**, 113-149, (1984).
129 Andreoni, J. Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* **37**, 291-304, (1988).
130 Henrich, J. *et al.* Markets, religion, community size, and the evolution of fairness and punishment. *science* **327**, 1480-1484, (2010).
131 Alesina, A., Giuliano, P. & Nunn, N. On the Origins of Gender Roles: Women and the Plough. *Quarterly Journal of Economics* **128**, 469-530, (2013).

132 Ellickson, R. C. in *Social Norms* eds Michael Hechter & Karl D Opp) 35-75 (Russell Sage Foundation, 2001).
133 Lowes, S., Nunn, N., Robinson, J. A. & Weigel, J. L. The evolution of culture and institutions: Evidence from the Kuba Kingdom. *Econometrica* **85**, 1065-1091, (2017).
134 Benabou, R. & Tirole, J. Laws and Norms. *working paper*, (2011).
135 Posner, E. A. *Law and Social Norms*. (Harvard University Press, 2000).
136 Sunstein, C. R. On the expressive function of law. *University of Pennsylvania Law Review* **144**, 2021-2053, (1996).
137 Akerlof, G. A. The missing motivation in macroeconomics. *American Economic Review* **97**, 5-36, (2007).
138 Allcott, H. Social norms and energy conservation. *Journal of Public Economics* **95**, 1082-1095, (2011).
139 Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J. & Griskevicius, V. Normative social influence is underdetected. *Personality and Social Psychology Bulletin* **34**, 913-923, (2008).

TEXT BOX ON IMPORTANT UNSOLVED RESEARCH PROBLEMS

There are still substantial gaps in our knowledge of the origins, the determinants and the consequences of social norms. Below we list 10 important questions and, if possible, add references that deal with these questions.

1) What are micro-sociological and psychological processes that facilitate and hinder the development of a social norm?
2) What is – at the conceptual level – the precise relationship between social preferences and social norms and how can we distinguish them empirically? How do social norms influence the motivational content of social preferences and, for given social preferences, how do they affect compliance with normative standards?
3) What determines individuals' agreement with the "ought component" of norms?[92] How do they come to internalize or reject a normative standard?
4) What explains the formation and the decay of social norms and how can we explain changes in the normative content, i.e., the "ought component" of social norms?[92]
5) What are the long-run environmental and economic determinants of social norms?[130-133] and how do normative standards evolve in the context of conflicting economic interests?[75]
6) How do economic incentives, the human desire for social approval and normative standards interact? When are they complements and when do economic incentives undermine normative standards and approval incentives?[134]
7) How does actual compliance and non-compliance shape the development of normative standards?[92]
8) Through which interventions and public policies is it possible to shape social norms[76] and which aspect of the norm and norm-related behaviors – the content of the normative standard, social agreement with the normative standard, behavioral compliance with the standard – is changed by the intervention?
9) How do legal institutions – apart from their sanctioning capacity – affect social norms and how do social norms affect the effectiveness of legal institutions?[134,135] To what extent do legal institutions shape normative standards by setting precedent, fall back rules or through expressing what is normatively approved and expected?[136]
10) To what extent and in which ways do social norms influence important economic and social patterns?[91,92,137-139]

Box on: How social preferences transform the prisoners' dilemma

In the prisoners' dilemma (PD) each of 2 players makes one of 2 choices: cooperate or defect. Material payoffs are such that regardless of the opponent's choice it is always in the self-interest of a player to defect (Table 1a). However, if both players defect they are worse off than if both cooperate, hence the dilemma.

Table 1a. Representation of prisoners' dilemma in terms of material payoffs

	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	0, 5
Defect (D)	5, 0	1, 1

If both players are sufficiently inequity averse their subjective preferences transform the PD into a different game in which mutual cooperation is also an equilibrium. In general, what subjects perceive as inequitable or equitable is likely to be shaped by society and the prevailing social norms[63] but the simplest form of inequity aversion is inequality aversion. In case of inequality aversion a player suffers from receiving less than the other with parameter α_i (envy), and also from receiving more than the other with parameter β_i (compassion). An inequality averse player i's subjective payoff u_i is a function of her own economic payoff x_i and of the payoff differences $(x_j - x_i)$ between the two players: $u_i = x_i - \alpha_i(x_j - x_i)$ if player i is worse off than player j $(x_j - x_i \geq 0)$, and $u_i = x_i - \beta_i(x_i - x_j)$ if player i is better off than player j $(x_i - x_j \geq 0)$. Inequality aversion makes unilateral defection less attractive by reducing the subjective payoff from 5 to $(5 - 5\beta)$ while being the victim of the other player's unilateral defection is particularly painful because it reduces the subjective payoff from 0 to -5α. Thus, for $\alpha = 1$ and $\beta = 0.5$ the PD is transformed into a coordination game in which both (C, C) and (D, D) are an equilibrium (Table 1b). Thus, if both players believe that the other one cooperates it is in their subjective interest to cooperate as well because the payoff of mutual cooperation is 4 while the payoff of unilateral defection is only 2.5.

Table 1b. Utility representation of prisoners' dilemma if players are inequality averse with parameters $\alpha = 1$ and $\beta = 0.5$

	Cooperate (C)	Defect (D)
Cooperate (C)	4, 4	–5, 2.5
Defect (D)	2.5, –5	1, 1

Similar results are obtained if both players have a sufficiently strong **preference for reciprocity** which can be modelled by a utility function such as $u_i = x_i + \gamma k_i k_{ij}^e$ where γ is positive and measures the strength of i's preferences for reciprocity, k_i is i's actual kindness to the other player, k_{ij}^e and is i's expected kindness of player j towards himself. $k_i > 0$ if player i is kind to player j whereas $k_i < 0$ if player i is unkind to player j. k_j is i's perceived kindness of player j. What is perceived as kind and unkind is likely to be shaped by social norms prevailing in society but it seems very plausible that cooperation in the PD is perceived as kind. Thus, suppose that each of the two players believes that the other one cooperates and perceives this to be kind ($k_{ij}^e > 0$). This means that each player can increase utility u_i

by reciprocating, i.e. being actually kind to the other player by cooperating ($k_i > 0$). Thus, as in the case of inequality aversion the material payoffs no longer represent the true utilities and for a sufficiently large γ both players derive more utility from mutual cooperation than from unilateral defection. However, if the player's believe that the other player is unkind by defecting ($k_{ij}^e < 0$) they have again an incentive to reciprocate by setting $k_i < 0$ through defection.

A similar albeit slightly different logic applies in the case of guilt averse individuals. Here, i's utility is given by $u_i = x_i - \gamma \max\{0,\ b_i^e - c_i\}$ where $\gamma > 0$ measures the propensity to feel guilty if one does not live up to what i believes that the other player expects from i, and $c_i \in \{0, \overline{c}\}$ denotes defection ($c_i = 0$) or cooperation ($c_i = \overline{c}$) . I's belief about the partner's expectation is denoted by b_i^e. Thus, if $b_i^e = \overline{c}$ and player i defects (i.e., $c_i = 0$), utility is given by $u_i = x_i - \gamma b_i^e$ which is – for sufficiently large guilt aversion smaller than γ – smaller than the utility of meeting the believed expectation of the other player.